What Does Deep Learning See? Insights From a Classifier Trained to Predict Contrast Enhancement Phase From CT Images

Kenneth A. Philbrick[1]
Kotaro Yoshida
Dai Inoue
Zeynettin Akkus
Timothy L. Kline
Alexander D. Weston
Panagiotis Korfiatis
Naoki Takahashi
Bradley J. Erickson

OBJECTIVE. Deep learning has shown great promise for improving medical image classification tasks. However, knowing what aspects of an image the deep learning system uses or, in a manner of speaking, sees to make its prediction is difficult.

MATERIALS AND METHODS. Within a radiologic imaging context, we investigated the utility of methods designed to identify features within images on which deep learning activates. In this study, we developed a classifier to identify contrast enhancement phase from whole-slice CT data. We then used this classifier as an easily interpretable system to explore the utility of class activation map (CAMs), gradient-weighted class activation maps (Grad-CAMs), saliency maps, guided backpropagation maps, and the saliency activation map, a novel map reported here, to identify image features the model used when performing prediction.

RESULTS. All techniques identified voxels within imaging that the classifier used. SAMs had greater specificity than did guided backpropagation maps, CAMs, and Grad-CAMs at identifying voxels within imaging that the model used to perform prediction. At shallow network layers, SAMs had greater specificity than Grad-CAMs at identifying input voxels that the layers within the model used to perform prediction.

CONCLUSION. As a whole, voxel-level visualizations and visualizations of the imaging features that activate shallow network layers are powerful techniques to identify features that deep learning models use when performing prediction.

Keywords: class activation map (CAM), computer-aided diagnosis, contrast enhancement phase, convolutional neural network (CNN), CT, deep learning, gradient-weighted class activation map (Grad-CAM), guided backpropagation, machine learning, saliency activation map, saliency map

doi.org/10.2214/AJR.18.20331

Received July 2, 2018; accepted after revision August 29, 2018.

Based on presentations at the 2018 Society for Imaging Informatics in Medicine annual meeting, San Francisco, CA, and the 2018 Radiological Society of North America annual meeting, Chicago, IL.

[1]All authors: Department of Radiology, Radiology Informatics Laboratory, Mayo Clinic, 3507 17th Ave NW, Rochester, MN 55901. Address correspondence to K. A. Philbrick (philbrick.kenneth@mayo.edu).

This article is available for credit.

Supplemental Data
Available online at www.ajronline.org.

AJR 2018; 211:1184–1193

0361–803X/18/2116–1184

The application of deep learning to perform radiologic diagnosis has gained much attention. Deep learning has shown great promise for doing many diagnostic tasks, but knowing what aspects of the image are used in making a decision is difficult [1–3]. In this study, we examined a rather simple deep learning task of recognizing the contrast enhancement phase of CT imaging as an example task and then attempted to discern the features the deep learning model used to perform image classification.

In contrast with traditional machine learning that relies on human-designed, precomputed features, deep learning classifiers categorize images by learning to recognize features contained within them [4, 5]. The ability to self-identify predictive features enables deep learning classifiers to learn both known and previously unrecognized predictive features in images. Identifying the image features used by the classifier to perform prediction provides a means to understand the classifier's decision-making process and to perform discovery science by identifying previously unrecognized predictive features in images.

The features identified by deep learning classifiers are difficult to describe directly. Deep learning classifiers learn by optimizing a series of interdependent nonlinear functions. For a given multilayer network, multiple visually distinct inputs can strongly activate a network's output. Techniques such as class-activation maps (CAMs), gradient-weighted class-activation maps (Grad-CAMs), saliency maps, and guided backpropagation maps have been developed to identify the features within an input image that a trained deep learning model identifies as being predictive [6–9].

IV contrast agents are routinely used in CT. After administration, contrast agent increases vascular and tissue attenuation (enhancement) in a time- and tissue-specific manner. Images are commonly acquired before and after contrast agent administration to gain information available at each phase, but physiologic variability, such as cardiac function, means that postinjection time is often not an accurate indicator of scan phase [10–12]. Instead, radiologists determine the scan phase from enhancement of normal

structures, but this process is subjective and may not be reproducible.

We developed a highly predictive deep learning classifier to identify the contrast enhancement phase of whole-slice CT data. We used knowledge of contrast phase to investigate imaging features and their locations that were identified by the deep learning model using CAMs, Grad-CAMs, saliency maps, guided backpropagation maps, and a novel map, the saliency-activation map, to assess the utility of these methods for identifying imaging features the deep learning model used in prediction. The results show that voxel-level visualizations provide powerful insights into the precise anatomic regions of the imaging that activated the network. These maps illustrate the regions of the image that the deep neural network uses when it makes its decision.

Materials and Methods

Dataset

Unenhanced and contrast-enhanced abdominal CT images were collected from patient examinations (1102 patients, 3253 examinations) conducted at the Mayo Clinic from 2001 to 2009. The study was approved by our institutional review board. Patients were imaged at one to five unique scan phases (mean ± SD, 2 ± 1 phases). The dataset was assigned by one of three radiologists to one of five contrast phases (unenhanced [noncon], corticomedullary [CM], late-CM to early nephrographic [CM–neph], nephrographic [neph], and pyelographic [pyelo]). They also identified a slice located near the middle of both kidneys. Dataset annotation was performed using custom software. Radiologists determined the scan phase on the basis of the enhancement of the kidneys and other structures such as liver and vessels. After the contrast enhancement phase was determined for all cases, multiple thumbnail images at the level of the kidneys of the same scan phase were displayed simultaneously, and any potentially mislabeled cases were identified by comparing with other thumbnail images. These cases were then reviewed again to reduce inter- and intraobserver variability. For each CT scan volume, images from the top to the bottom of the kidneys were extracted. Examinations (n = 533) with multiple reconstructions of the same scan phase were included to allow the effects of differences in CT reconstruction parameters to be learned by the classifier. Patient images were randomized and divided into training (n = 705), validation (n = 176, 20% of [training + validation]), and test (n = 221; 20% of total dataset) datasets. Datasets were randomized once, and the same training, validation, and test datasets were used in all model experiments.

Data Standardization and Augmentation

Voxels outside the body were removed, the mean value was subtracted from the images (mean = 0), and the values were rescaled to unit SD (training dataset mean attenuation ± SD, –23.3 ± 144.1 HU). Input training data were augmented using random rotation ± 5° in slice rotation and random ± 5 voxels in slice translation.

Models

Categoric prediction convolutional neural networks (CNNs) were built to predict scan phase from any single slice or three consecutive images of input data [13]. Models were built using Keras Deep Learning library with a TensorFlow backend engine (version 1.8.0, TensorFlow). Models were loosely based on VGG architecture for simplicity and clarity purposes. Block diagrams of the architectures tested are shown in Figure S1 in the *AJR* electronic supplement to this article (available at www.ajronline.org). All convolutional layers used rectified linear unit activation functions [14]. Grid search was used to optimize model architecture. Models were trained using the Adam optimizer, with learning rate initialized to 0.0001 and a learning rate decay of 0.316 after the validation loss failed to improve for four consecutive epochs [15]. Models were trained until the validation loss failed to improve for 12 consecutive epochs. Balanced class weighting was used to account for differences in input slice number between the scan phases and between patients within a phase. The top performing model selected for subsequent visualization experiments accepted whole slice data (512 × 512 voxels) and processed the data through five max pooled blocks of layers, composed of two 3 × 3 convolutions and a max pooling layer, to produce final kernel output of 16 × 16 × 64. Also, for comparison purposes, reference implementations of VGG16, InceptionV3, Resnet50, DenseNet201, and InceptionResNet were trained [16–20].

Scan Phase Prediction

After model training, scan phase predictions were performed against the test dataset using both single-slice prediction and multislice voting. For single-slice prediction, model predictions were made directly for each available slice of data. Multislice voting was based on the premise that images within a scan volume would differ in their power to predict contrast enhancement phase and that the model's prediction could be improved by combining predictions from multiple images. We implemented Pandemonium voting and used the winning contrast enhancement phase probability to weight the slice prediction vote [21, 22].

Visualization of the Imaging Features Identified by Deep Learning

Saliency maps, CAMs, Grad-CAMs, guided backpropagation maps, and saliency activation maps were generated for the best performing model against the test input dataset. In brief, CAMs illustrate the spatial activation of the final output layer of a CNN with respect to a specific network output. CAMs are computed as the weighted sum of the final CNN layer kernels' output [9]. Saliency maps illustrate the voxels within an image that the model would alter to improve the model predictions of a specific output. Saliency maps are computed as the derivative of input imaging with respect to a specific network output [7]. Grad-CAMs illustrate the relative positive activation of a convolutional layer with respect to network output. Grad-CAMs are computed as the rectified linear unit of the gradient weighted sum of a layer's output [6]. Guided backpropagation maps illustrate voxels within the input that contribute positively to predict a specific output. Guided backpropagation maps are created by selectively backpropagating the positive component of the gradient calculated between input data and a network output [8].

Saliency-activation maps are a novel visualization designed to address limitations of Grad-CAMs. Saliency activation maps were designed to capture changes in the gradient across the CNN layer kernels and to capture network activations in regions with very low negative gradients and near-zero kernel output.

We used the following equation to calculate the saliency activation map:

$$L^{t}_{x,y,i} = \begin{cases} \max(L_{:,:,i}) - L_{x,y,i},\ G(K_{x,y,i}) < 0 \\ L_{x,y,i},\ G(K_{x,y,i}) \geq 0 \end{cases},$$

$$\sum_{x,y,i} | G(K_{x,y,i}) | \cdot L^{t}_{x,y,i},$$

where K is kernel gradient; L is kernel output; L^t is transformed kernel output; x,y is elementwise position within the kernel gradient and output; G is gaussian blur function ($\sigma = 0.75$); and i is layer kernel index.

Graphical Visualization

Saliency maps are visualized as the rank order visualization of the absolute value of the map. Guided backpropagation maps are visualized as the rank order visualization of the map. CAM, saliency activation map, and Grad-CAM are direct visualizations of the values in the map. Contrast enhancement phase was correctly predicted for shown images. Heat map visualizations were shown relative to the range of values within the image. Values in the bottom 2.5% of the map were not shown for clarity. All visualizations used a rainbow color map projection, in which red signified high attenuation and purple signified low attenuation.

Quantification of Visualization Specificity

Voxels within images that were correctly classified by the model were selectively masked to quantify the specificity of the voxels identified by each of the investigated techniques [23]. Specifically, values represented in the maps were ranked by absolute value. Imaging voxels corresponding with the upper 10% of the map or lower 10% (control comparison) were overwritten with random values to mask them. The probability of the correct contrast enhancement phase classification was then calculated for the masked image and normalized by the probability calculated for the original imaging, that is, (100 × masked probability) / original imaging probability. Values less than 100 provide evidence that the masking removed voxels from the image that the model used to perform prediction.

Statistics

The validation dataset was used to guide hyperparameter optimization and model selection. Performance metrics were computed on the test dataset after final model selection. Confusion matrix, precision, recall, F1 score, ROC AUC, and ROC curves (one vs all) were created for the top performing model. We used *t* tests to assess differences between single-slice and multislice voting prediction and to determine if targeted masking affected the model's prediction. A Bonferroni correction for multiple comparisons was applied where appropriate. Statistical significance was set at $p < 0.05$.

Results

Contrast Enhancement Phase Prediction Model Selection

Twenty-seven CNNs were trained to predict contrast enhancement phase from CT to optimize network hyperparameters. Precision, recall, F1 score, and ROC AUC (one vs all) metrics for all models tested are shown in Tables S2 and S3 (which can be viewed in the *AJR* electronic supplement to this article, available at www.ajronline.org). The model with the highest ROC AUC on the validation dataset using single-slice prediction and multislice voting was selected as the top performing model. Using single-slice CNN prediction, this model exhibited precision, recall, F1 score, and ROC AUC metrics of 0.776, 0.862, 0.860, and 0.981 on the test dataset, respectively. Using multislice voting, the model exhibited precision, recall, F1 score, and ROC AUC of 0.869, 0.923, 0.921, and 0.988, respectively. Figure S4 (which can be viewed in the *AJR* electronic supplement to this article, available at www.ajronline.org) shows training loss and accuracy curves and one-versus-all ROC curves for the top performing model. Table 1 gives confusion matrixes generated against the test dataset, and Table 2 gives precision, recall, F1 score, and ROC AUC metrics generated against the test dataset.

After model optimization and selection, test dataset performance metrics were computed for all tested models. No differences were detected ($p > 0.05$) between the validation and test dataset performance metrics using either single-slice prediction or multislice voting. Multislice voting increased model precision on the validation and test datasets ($p < 0.05$); it exhibited trends for increased recall on the validation and test datasets and for F1 score on the test dataset ($p < 0.1$).

TABLE 1: Confusion Matrixes of the Top Performing Model When Predicting Contrast Enhancement Phase for the Test Dataset

Predicted Phase	Labeled Phase				
	Noncon	CM	CM–Neph	Neph	Pyelo
Single-slice prediction only					
Noncon	2876	12	1	25	36
CM	7	2247	179	13	13
CM–neph	0	216	1199	223	14
Neph	2	16	292	1548	144
Pyelo	0	8	65	153	2876
Single-slice prediction followed by multislice voting					
Noncon	141	0	0	0	0
CM	0	102	7	0	0
CM–neph	0	7	61	8	0
Neph	0	0	9	64	3
Pyelo	0	0	3	2	119

Note—Noncon = unenhanced, CM = corticomedullary, CM–neph = corticomedullary to early nephrographic, Neph = nephrographic, Pyelo = pyelographic.

TABLE 2: Classification Performance Metrics for Convolutional Neural Network Single-Slice Classification and Single-Slice Classification Followed by Multislice Voting on the Test Dataset

Model	Precision		Recall		F1 Score	ROC AUC
	Overall	Within One Phase	Overall	Within One Phase		
Single-slice prediction only						
Noncon	0.975	0.991	0.997	0.999	0.986	0.999
CM	0.914	0.989	0.899	0.990	0.906	0.993
CM–neph	0.726	0.912	0.691	0.962	0.691	0.946
Neph	0.773	0.991	0.789	0.981	0.781	0.969
Pyelo	0.927	0.976	0.933	0.991	0.930	0.991
Single-slice prediction followed by multislice voting						
Noncon	1.000	1.000	1.000	1.000	1.000	1.000
CM	0.938	1.000	0.936	1.000	0.950	0.997
CM–neph	0.851	0.963	0.763	0.963	0.818	0.964
Neph	0.878	1.000	0.865	1.000	0.878	0.985
Pyelo	0.960	1.000	0.975	1.000	0.972	0.996

Note—Noncon = unenhanced, CM = corticomedullary, CM–neph = corticomedullary nephrographic, Neph = nephrographic, Pyelo = pyelographic.

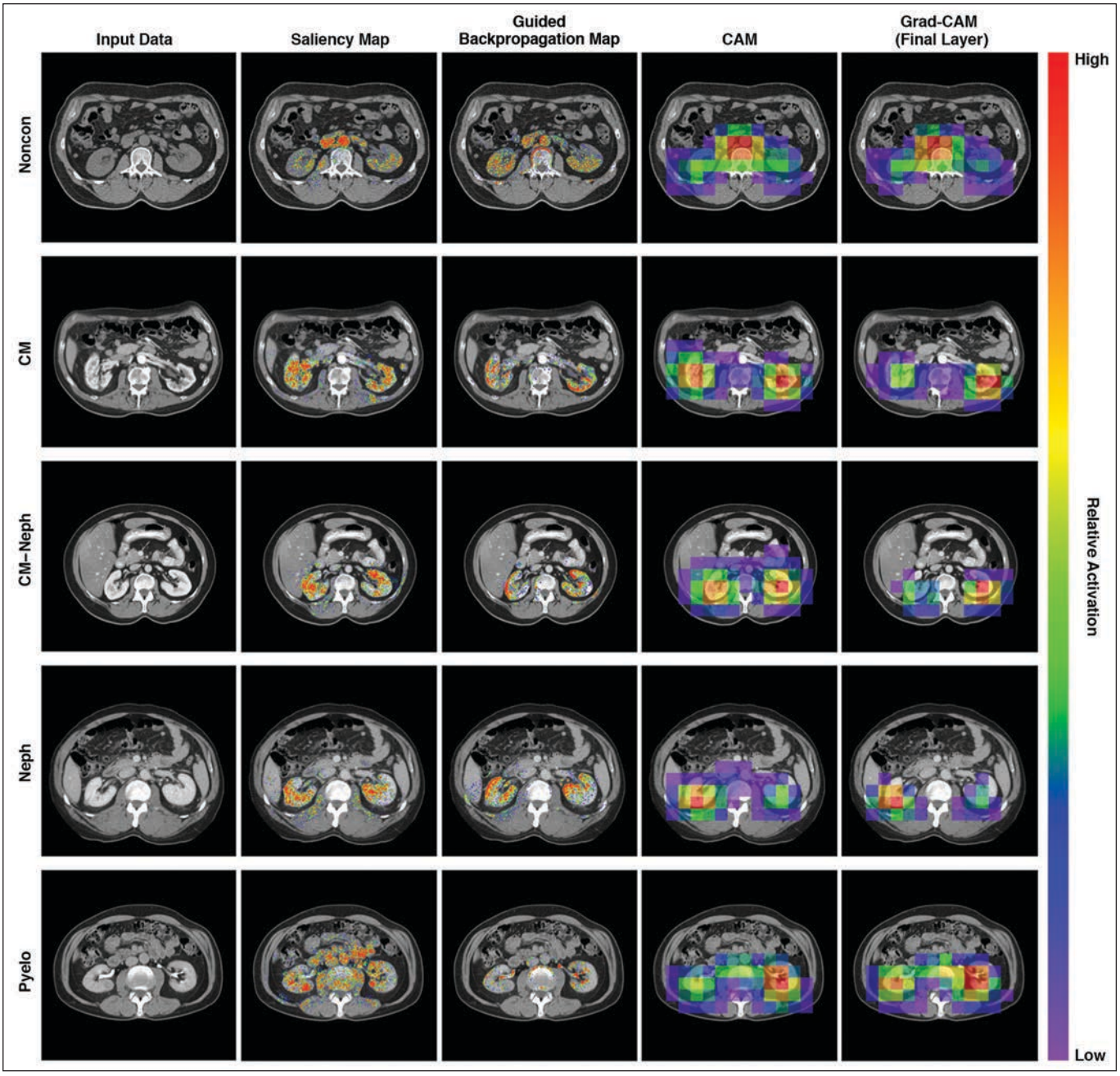

Fig. 1—Representative saliency maps, guided backpropagation maps, class activation maps (CAMs), and gradient-weighted class activation maps (Grad-CAMs) (final layer) for test dataset imaging. Noncon = unenhanced, CM = corticomedullary, CM–neph = corticomedullary to early nephrographic, Neph = nephrographic, Pyelo = pyelographic.

Reference implementations of VGG16, ResNet50, InceptionV3, InceptionResNet, and DenseNet201 were trained to predict contrast enhancement phase for comparison purposes. Figure S5 (which can be viewed in the *AJR* electronic supplement to this article, available at www.ajronline.org) shows training and validation loss curves and one-versus-all ROC curves computed against the test dataset. The VGG16 model failed to learn. Validation loss for all other models was minimized within the first three training epochs. None of the models exhibited average ROC AUC greater than the model described in this article. Table S6 (which can be viewed in the *AJR* electronic supplement to this article, available at www.ajronline.org) shows confusion matrixes generated using single-slice prediction against the test dataset.

Image Features on Which the Deep Learning Model Activates

Saliency maps and guided backpropagation maps broadly identify the relative importance that voxels in the input imaging had on the model's scan phase class prediction. Figure 1 shows saliency maps and guided backpropagation maps computed for all correctly classified test dataset imaging and

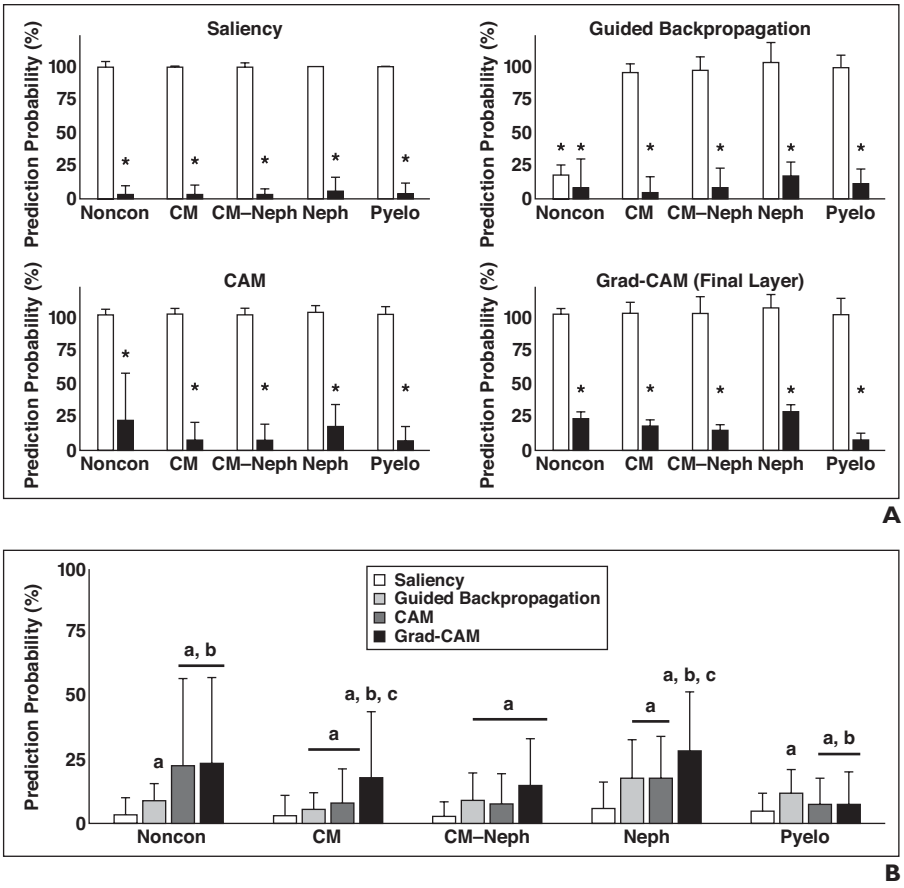

Fig. 2—Effect of map-targeted image masking on model prediction probability. Noncon = unenhanced, CM = corticomedullary, CM–neph = corticomedullary to early nephrographic, Neph = nephrographic, Pyelo = pyelographic.
A, Graphs show masking upper 10% of voxels (*black bars*) identified in saliency map, guided backpropagation map, class activation map CAM, and gradient-weighted class activation map (Grad-CAM) reduced model's prediction probability for correct contrast enhancement phase. Masking bottom 10% of voxels identified by maps (*white bars*) did not alter contrast enhancement phase prediction probability. Data are means with whiskers representing SD. Asterisks indicate statistically significant difference from 100% ($p < 0.05$).
B, Relative specificity of visualizations. Saliency map had greatest specificity for voxels that, when masked, lowered convolutional neural network prediction for correct phase. Lower bars indicate selective voxel masking exerted greater effect to attenuate model's slice prediction probability for correct phase. Data are means with whiskers representing SD. a = different from saliency map, b = different from guided backpropagation map, c = different from CAM; ($p < 0.05$).

representative images. Qualitatively, saliency maps identified voxels associated with the kidney, renal vasculature, aorta, vena cava, and to a lesser extent muscle, liver, and, in the pyelographic phase, the digestive tract as being predictive. Guided backpropagation maps identified voxels in the kidney, renal vasculature, aorta, vena cava, and to a lesser extent muscle and bone as being predictive. As a whole, the visualizations generated by the guided backpropagation identified fewer voxels and more clearly identified the renal cortex as being an anatomic region that predicted contrast enhancement phase.

Saliency maps represent the relative change, both positive and negative, that the model predicts would improve the prediction for a specific classification. The map shown in Figure 1 illustrates the absolute value of the saliency map to emphasize the regions of the map in which the model would make the greatest change irrespective of direction. Figure S7 (which can be viewed in the *AJR* electronic supplement to this article, available at www.ajronline.org) illustrates the same map paired with a map illustrating the rank order signed value contained within the saliency map. Regions that are high in the absolute value representation correspond with regions of the extreme high and extreme low in the value representation. The value representation illustrates that the model expects unenhanced imaging to exhibit low attenuation values within the aorta and vena cava and CM phase imaging to exhibit relatively high attenuation values within the renal cortex and low values within the medulla.

CAMs and Grad-CAMs, computed against the last CNN layer, illustrate the relative importance of the spatial regions described in the CNN final output layer on the model's prediction. These visualizations generate lower resolution maps and thereby indirectly identify blocks of input voxels. CAMs and Grad-CAMs were computed for all correctly classified test set imaging; Figure 1 shows representative images. Qualitatively, these maps identified regions of voxels that correspond with the kidney, renal vasculature, aorta, vena cava, bone, and spinal muscles. The low resolution of these maps makes it impossible to precisely identify voxel-level features in the input imaging on which the network activated.

Masking the most influential regions identified in each of the maps (top 10% of the map) greatly lowered the model's predicted probability for the slice's correct contrast enhancement phase ($p < 0.05$) (Fig. 2A). In contrast, masking the least influential regions identified in each of the maps (bottom 10% of the map) had little to no effect on classifier prediction (Fig. 2A).

The specificity of techniques was not equal (Fig. 2B). Saliency maps provided greater specificity for voxels in the input imaging, which affected scan phase prediction more than the other techniques. For some, but not all, scan phase predictions, CAMs exhibited slightly greater specificity than Grad-CAMs (computed for the final layer).

Visualization of Intermediate Network Layers

The method used to generate Grad-CAMs has been suggested to be applicable to visualize the imaging features that activate intermediate network layers [6]. Figure 3 shows representative Grad-CAMs generated for CNN network layers immediately preceding the model's max pooling layers. The visualizations of network activations at layer 5b and to a lesser extent layer 4b, near the final network layer (layer 6), appear to correspond

with input image regions that are predictive of contrast enhancement phase. In contrast, visualizations of shallower network layers (layers 2b and 3b) appear to exhibit little correspondence with regions of the input imaging that are logically predictive of contrast enhancement phase.

We developed a novel visualization, saliency activation maps, to generate improved activation maps of the intermediate network layers. Figure 4 illustrates representative saliency activation map output for network layers 2b, 3b, 4b, 5b, and 6 for the same imaging shown in Figure 3 for Grad-CAM. Qualitatively, at shallower network layers (layers 2b, 3b, and 4b) saliency activation map clearly identified regions within the kidney, renal vasculature, aorta, vena cava, and to a much lesser extent within the surrounding tissues as affecting the model's scan phase prediction.

Across all scan phases the voxels identified in saliency activation maps exhibited greater specificity for scan phase prediction than those identified by Grad-CAMs for layers 2b, 3b, and 4b (excluding unenhanced imaging) ($p < 0.05$) (Fig. 5). We found no difference in the specificity of saliency activation maps and Grad-CAMs at layer 5b ($p > 0.05$) except for unenhanced scan phase prediction. At layer 6, we found no difference in specificity between saliency activation maps and Grad-CAMs for any scan phase.

Quantification of the relative specificity of the voxels identified in saliency activation maps at model layers 2b, 3b, 4b, 5b, and 6 on model contrast enhancement phase prediction illustrate that voxels identified by saliency activation maps at shallower layers (2b, 3b, and 4b) exhibited greater specificity than voxels identified at deep layers (5b and 6) (Fig. 6).

Discussion

CNNs broadly identify features by activating on relatively simple patterns and then repeatedly combining these patterns to identify complex shapes and textures across broad

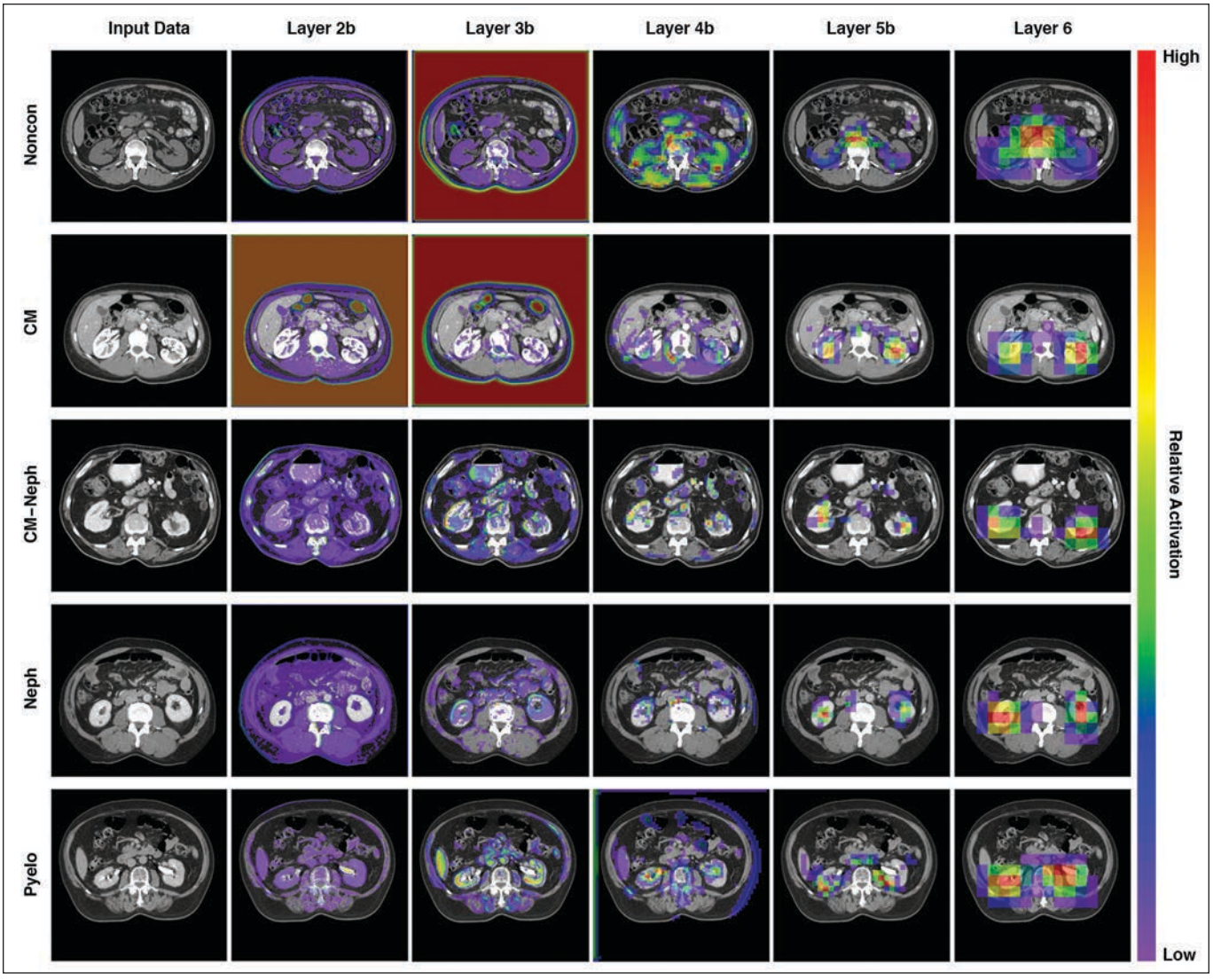

Fig. 3—Representative gradient-weighted class activation maps generated for test dataset imaging. Maps from deep layers (5b and 6) spatially correspond with anatomic locations that reflect contrast enhancement phase (kidney and aorta). However, at shallow layers (2b, 3b, and 4b), maps do not selectively identify image regions that are predictive of contrast enhancement phase. Noncon = unenhanced, CM = corticomedullary, CM–neph = corticomedullary to early nephrographic, Neph = nephrographic, Pyelo = pyelographic.

spatial areas. We trained a CNN classifier to identify contrast enhancement phase from whole-slice axial CT. CT images were classified into one of five phases: unenhanced (noncon), corticomedullary (CM), late CM to early nephrographic (CM–neph), nephrographic (neph), and pyelographic (pyelo). The model exhibited excellent performance. The preponderance of the model's errors resulted in misclassification to a phase immediately before or after the ground truth annotated phase. Misclassification was most common for data illustrating CM–neph and neph scan phases. Given that contrast enhancement phase is a discrete classification of a continuous event, this result suggests that the errors in the classification occurred primarily in discerning the transitions between the phases.

Ensemble methods combine the output of multiple classifiers, or as shown in this study, voting systems that combine multiple outputs from the same classifier are used to improve prediction performance after training [21, 24]. We trained the reported classifier to predict contrast enhancement phase from a single image. The scan phase prediction probabilities generated by the classifier varied over the sampled images. Logically, the predicted probabilities were lower for images that contained imaging that the classifier identified as being predictive of multiple contrast enhancement phases. We reasoned, given that the predictive power of individual images was not equal, the overall phase prediction for the volume as a whole could be improved using a Pandemonium style voting scheme and weighting a volume's slice votes by the slice's phase prediction probability [21, 22]. This system improved the precision of all models tested and exhibited a trend toward improved recall and F1 score. Though beyond the scope of this work, this finding suggests that a 3D classification model should perform at least as well as the voting system we report and likely better because a 3D model could perform more sophisticated multislice prediction and add cross-slice spatial information to the classification model.

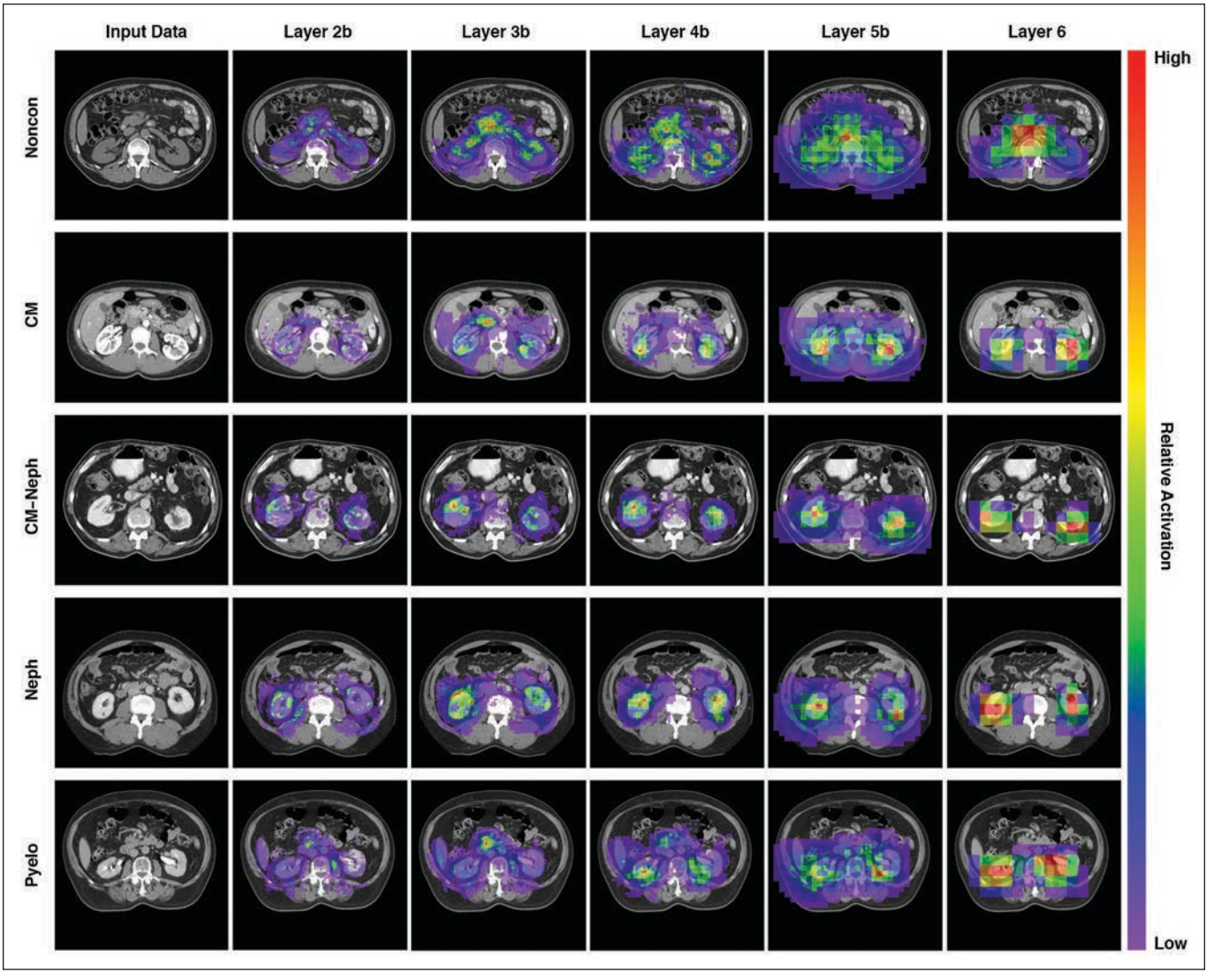

Fig. 4—Representative saliency activation maps generated for test dataset imaging. Maps at all layers spatially correspond with anatomic locations that reflect contrast enhancement phase. Noncon = unenhanced, CM = corticomedullary, CM–neph = corticomedullary to early nephrographic, Neph = nephrographic, Pyelo = pyelographic.

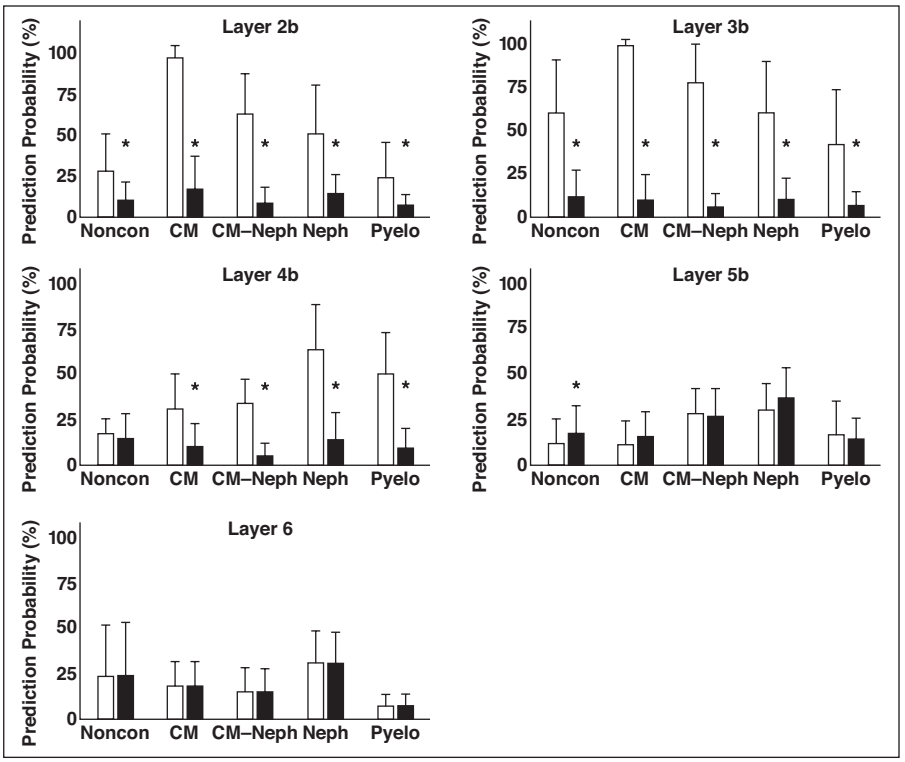

Fig. 5—Saliency activation maps (SAMs) (*black bars*) had greater specificity than gradient-weighted class activation maps (Grad-CAM, *white bars*) at shallow network layers (2b, 3b, and 4b) for voxels that, when masked, lowered convolutional neural network prediction for correct contrast enhancement phase. Data are means with whiskers representing SD. Asterisks indicate SAM difference from Grad-CAM ($p < 0.05$). Noncon = unenhanced, CM = corticomedullary, CM–neph = corticomedullary to early nephrographic, Neph = nephrographic, Pyelo = pyelographic.

The primary purpose of this study was to develop a well-performing classifier through which to investigate methods that visualize the features identified by CNN classifiers. A number of CNN architectures (VGG, Inception, ResNet, DenseNet) have been developed to perform image classification. These architectures have been developed in reference to the ImageNet classification challenge [25]. We chose to base our model on VGG architecture for simplicity and clarity purposes. In total, the model reported here contained 132,181 trainable parameters, far fewer than VGG16 (~138 million), InceptionV3 (~23.8 million), ResNet50 (~25.6 million), InceptionResNet (~55.8 million), and DenseNet201 (~20.2 million).

After developing our classifier, we trained and tested reference models of VGG16, InceptionV3, ResNet50, InceptionResNet, and DenseNet201 [16–20]. The failure of VGG16 to learn on our training dataset was likely a consequence of the model's enormous parameter size and the absence of architectural features that mitigate the vanishing gradient problem associated with training very deep neural networks [26]. Though trainable, when trained on our dataset InceptionV3, ResNet50, InceptionResNet, and DenseNet201 all rapidly overfit our training dataset. This finding was not surprising because these networks have been optimized to solve the ImageNet challenge, a much larger classification challenge than the problem we posed [25]. Correspondingly, the trainable parameter space of InceptionV3, ResNet50, InceptionResNet, and DenseNet201 are much larger than in our model. Our network, with its much lower parameter space, forced the model to converge to a solution that fit the training dataset less tightly. As a result, our model required more training epochs to converge to a solution that minimized the loss on the validation dataset. The greater performance metrics exhibited by our model than the reference models provides evidence that our simpler model generalized better to the dataset as a whole. However, it is entrely possible that with additional optimization architectures inspired by the designs of Inception, Resnet, or Densenet could have achieved superior performance metrics than the model we report.

Saliency maps, guided backpropagation maps, CAMs, and Grad-CAMs have been proposed as methods to identify imaging features that activate a CNN [6–9]. To explore the utility of these methods within a radiologic imaging context, we generated these visualizations for the reported contrast enhancement phase classifier. All methods selectively identified voxels within input imaging that activated the network. However, the specificity of the methods was not equal. As a whole, saliency maps in particular exhibited greater specificity for input voxels that the model used to perform prediction.

The high specificity of saliency maps for predictive features within the imaging is perhaps not surprising because this method directly identifies voxels within input imaging. Saliency maps are computed as the gradient of the input voxels with respect to the network output [7]. Intuitively, saliency maps thereby represent the relative change, both positive and negative, the network would make to its input to increase the prediction probability for a specific class. Visualizations of both the saliency map absolute value and relative value provide information about a classifier's response to its input. Maps of the saliency map's absolute value identify areas of image the classifier predicts would exert the greatest effect on its prediction if they were changed. As shown, selectively altering the voxels within these regions greatly attenuates the classifier's predictions. Visualizations of the saliency map relative value provide information on the relative direction that the model suggests for changing the identified voxel's values (addition or subtraction) to improve its prediction. These gradient images are typically noisy [27]. As we have shown, qualitatively these visualizations can be used to infer relative values the model anticipates finding within regions of an image for a given classification.

Guided backpropagation maps are produced by selectively backpropagating the positive component of the gradient [8]. Guided backpropagation maps are designed to selec-

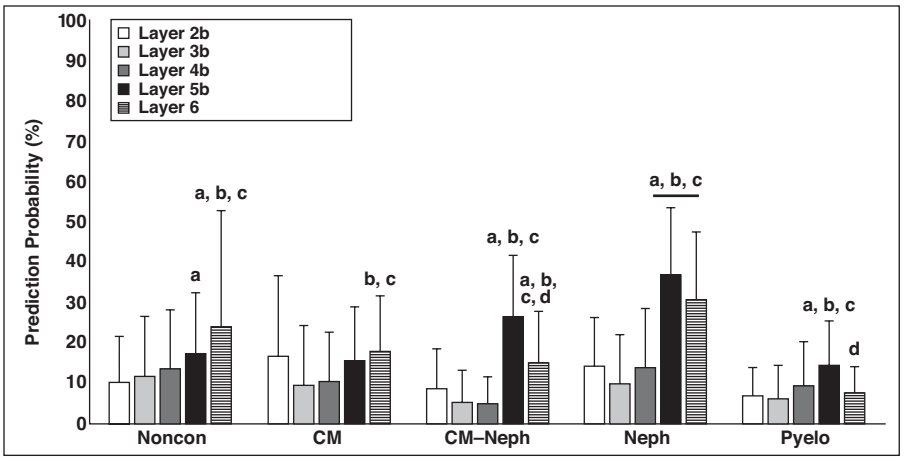

Fig. 6—Saliency activation map specificity for voxels that, when masked, lowered convolutional neural network prediction probability for correct contrast enhancement phase and was greater at shallower layers (2b, 3b, 4b). Lower bars indicate selective voxel masking exerted greater effect to attenuate model's slice prediction probability for correct phase. Data are means with whiskers representing SD. Noncon = unenhanced, CM = corticomedullary, CM–neph = corticomedullary to early nephrographic, Neph = nephrographic, Pyelo = pyelographic; a = different from layer 2b, b = different from layer 3b, c = different from layer 4b, d = different from layer 5b; $p < 0.05$.

tively identify imaging features that induce positive activations at all levels of the network. As would be expected, guided backpropagation maps identified a more focused region of the input than the saliency map. However, the lower specificity of the guided backpropagation map compared with the saliency map provides evidence that the model used regions of the input image that induced positive and negative model gradients.

CAMs and Grad-CAMs, computed against the last CNN layer, are weighted visualizations of the final CNN kernel output [9, 28]. For networks with small output dimensions, this characteristic may limit the utility of these methods. When laid over input imaging, these methods have been suggested to indirectly identify voxels to which the network responds. Experimentally, these methods have been used to enable the localization of features identified by CNN classifiers [9, 28, 29]. The specificity exhibited by CAMs and Grad-CAMs was lower than that exhibited by saliency maps and guided backpropagation maps. The lower resolution of these maps logically likely contributed to this reduction in specificity.

Convolutional kernels of size 3 × 3 or greater, and max pooling layers, act to move information horizontally across network layers, which enables the network to identify relationships between imaging features that are spatially distant in the input imaging [30]. However, for CAMs and Grad-CAMs the horizontal movement of information across the network also promotes dissociation between the map and the original input image. In the examples shown, the partial correspondence between blocks of high activity in CAMs and Grad-CAMs and anatomic structures (e.g., aorta) identified in the saliency and guided backpropagation maps provides evidence that the horizontal movement of information across layers lowers the voxel specificity of CAMs and Grad-CAMs.

In contrast with saliency maps, guided backpropagation maps, and CAMs, the algorithm used to generate Grad-CAMs has been suggested to be applicable to identify the features in imaging that activate any network layer, suggesting that Grad-CAMs could be used to identify features in imaging that activate intermediate network layers [6]. The ability to generate Grad-CAMs at layers closer to the input is logically advantageous. Kernel output at shallower layers has a higher resolution and has undergone less horizontal translation than at deeper layers as a result of having been processed through fewer convolutional and max pooling layers. However, in our data, Grad-CAMs generated for relatively shallow network layers (layers 2b and 3b) exhibited little correspondence with regions of the imaging that were predictive of contrast enhancement phase; in the worst cases, these maps identified nonimage background as being predictive of contrast enhancement phase.

We developed saliency activation maps as an alternative to the Grad-CAM. Saliency activation maps were designed to address two failings of the Grad-CAM: errors associated with kernel weighting and errors associated with the representation of low kernel output values.

Grad-CAMs weight kernel output by the mean of the kernel's gradient. This weighting scheme weights all voxels of a kernel's output equally. For the final layer in CNN, this weighting scheme correctly approximates the CAM. However, at intermediate network layers, mean gradient weighting can incorrectly emphasize kernel output for voxels that have a high output value but a low element-wise gradient. In the saliency activation map, we use the elementwise product of the kernel's output and the kernel's gradient to conduct element-wise instead of kernel-wide weighting. This weighting scheme allows the map to selectively emphasize subregions of a kernel's output.

Grad-CAM identifies the relative importance of a kernel as the product of the kernel's output and its weight. This weighting scheme assumes that large output values act to promote eventual network activation. However, for regions of a layer's kernel, a negative gradient indicates that smaller output values and not larger output values are desirable. As an example, if the output of a kernel is negatively weighted in a network layer, low kernel output values and not larger values will act to promote the subsequent network layer's activation (assuming rectified linear unit). To account for this, saliency activation maps selectively invert kernel output (max [kernel output] – kernel output) for kernel elements with a negative elementwise gradient weights.

Saliency activation maps computed at shallow layers (layers 2b and 3b) clearly identified anatomic structures that are affected by vascular contrast agents and exhibited greater voxel specificity than saliency activation maps from deeper layers. This finding supports the concept that the visualization of the imaging features identified by CNN within shallower network layers may provide more information than deeper network layers because the kernel output from shallow network layers has been processed through fewer convolutional layers and max pooling layers.

When utilizing deep learning models, understanding what deep learning sees is critical for trusting results and gaining insight into disease. In this study, we used CT scan phase as a paradigm to illustrate the ability to show both where deep learning is looking and the properties of the image that the network is activating on, or seeing, to perform classification. Our results indicate that, as a whole, voxel-level vi-

sualizations and visualizations of the imaging features that activate shallower network layers are particularly powerful techniques to identify the image features that deep learning models see when performing prediction.

References

1. Akkus Z, Ali I, Sedlář J, et al. Predicting deletion of chromosomal arms 1p/19q in low-grade gliomas from MR images using machine intelligence. *J Digit Imaging* 2017; 30:469–476
2. Kline TL, Korfiatis P, Edwards ME, et al. Performance of an artificial multi-observer deep neural network for fully automated segmentation of polycystic kidneys. *J Digit Imaging* 2017; 30:442–448
3. Korfiatis P, Kline TL, Lachance DH, Parney IF, Buckner JC, Erickson BJ. Residual deep convolutional neural network predicts MGMT methylation status. *J Digit Imaging* 2017; 30:622–628
4. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, eds. *Advances in neural information processing systems 25*. Red Hook, NY: Curran Associates, 2012:1097–1105
5. LeCun Y, Boser B, Denker JS, et al. Backpropagation applied to handwritten zip code recognition. *Neural Comput* 1989; 1:541–551
6. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. arXiv website. arxiv.org/pdf/1610.02391.pdf. Published October 7, 2016. Updated March 21, 2017. Accessed September 20, 2018
7. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv website. arxiv.org/pdf/1312.6034.pdf. Published December 20, 2013. Updated April 19, 2014. Accessed September 20, 2018
8. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for simplicity: the all convolutional net. arXiv website. arxiv.org/pdf/1412.6806.pdf. Published December 21, 2014. Updated April 13, 2015. Accessed September 20, 2018
9. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. arXiv website. arxiv.org/pdf/1512.04150.pdf. Published December 14, 2015. Accessed September 20, 2018
10. Bae KT, Heiken JP, Brink JA. Aortic and hepatic contrast medium enhancement at CT. Part II. Effect of reduced cardiac output in a porcine model. *Radiology* 1998; 207:657–662
11. Bae KT. Intravenous contrast medium administration and scan timing at CT: considerations and approaches. *Radiology* 2010; 256:32–61
12. Lawaczeck R, Jost G, Pietsch H. Pharmacokinetics of contrast media in humans: model with circulation, distribution, and renal excretion. *Invest Radiol* 2011; 46:576–585
13. Sahiner B, Chan HP, Petrick N, et al. Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. *IEEE Trans Med Imaging* 1996; 15:598–610
14. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. Haifa, Israel: Omnipress, 2010:807–814
15. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv website. arxiv.org/pdf/1412.6980.pdf. Published December 22, 2014. Updated January 30, 2017. Accessed September 26, 2018
16. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *2017 IEEE conference on computer vision and pattern recognition (CVPR)*. Piscataway, NJ: IEEE, 2017:2261–2269
17. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *2016 IEEE conference on computer vision and pattern recognition (CVPR)*. Piscataway, NJ: IEEE, 2016:770–778
18. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, Inception-ResNet and the impact of residual connections on learning. arXiv website. arxiv.org/pdf/1602.07261.pdf. Published February 23, 2016. Updated August 23, 2016. Accessed September 26, 2018
19. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: *2016 IEEE conference on computer vision and pattern recognition (CVPR)*. Piscataway, NJ: IEEE, 2016:2818–2826
20. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv website. arxiv.org/pdf/1409.1556.pdf. Published September 4, 2014. Updated April 10, 2015. Accessed September 26, 2018
21. van Erp M, Vuurpijl L, Schomaker L. An overview and comparison of voting methods for pattern recognition. In: *2002 proceedings eighth international workshop on frontiers in handwriting recognition*. Piscataway, NJ: IEEE, 2002:195–200
22. Selfridge OG. Pandemonium: a paradigm for learning. In: *Anderson JAD, Rosenfeld E, eds. Neurocomputing: foundations of research*. Cambridge, MA: MIT Press, 1988:115–122
23. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, eds. *Computer vision: ECCV 2014*. Cham, Switzerland: Springer International Publishing, 2014:818–833
24. Dietterich TG. *Ensemble methods in machine learning*. In: Kittler J, Roli F, eds. *Multiple classifier systems*. Berlin, Germany: Springer-Verlag, 2000:1–15
25. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vis* 2015; 115:211–252
26. Veit A, Wilber M, Belongie S. Residual networks behave like ensembles of relatively shallow networks. In: Lee DD, von Luxburg U, Garnett R, Sugiyama M, Guyon I, eds. *Advances in neural information processing systems 30*. Red Hook, NY: Curran Associates, 2016:550–558
27. Olah C, Mordvintsev A, Schubert L. Feature visualization. *Distill* 2017; 2:00007
28. Selvaraju RR, Das A, Vedantam R, Cogswell M, Parikh D, Batra D. Grad-CAM: why did you say that? arXiv website. arxiv.org/pdf/1611.07450.pdf. Published November 22, 2016. Updated January 25, 2017. Accessed September 26, 2018
29. Lu J, Xiong C, Parikh D, Socher R. Knowing when to look: adaptive attention via a visual sentinel for image captioning. In: *2017 IEEE conference on computer vision and pattern recognition (CVPR)*. Piscataway, NJ: IEEE, 2017:3242–3250
30. Luo W, Li Y, Urtasun R, Zemel R. Understanding the effective receptive field in deep convolutional neural networks. In: Lee DD, von Luxburg U, Garnett R, Sugiyama M, Guyon I, eds. *Advances in neural information processing systems 30*. Red Hook, NY: Curran Associates, 2016:4905–4913